Comparing Variance Partitioning and the Residual Method for Interpreting Brain Recordings

Leo Schultheiß (leo.schultheiss@tum.de)
Technical University Munich
Munich, Germany

Subba Reddy Oota (subba.reddy.oota@tu-berlin.de)
Technical University Berlin
Berlin, Germany

Anwar Nunez-Elizalde (anwarnunez@gmail.com)
Independent

Fatma Deniz (deniz@tu-berlin.de)
Technical University Berlin
Bernstein Center for Computational Neuroscience Berlin
Berlin, Germany

Abstract

A shift toward more naturalistic experiments in computational cognitive neuroscience has enabled a richer analysis of brain recordings. These naturalistic experiments often allow for the extraction of multiple feature spaces from stimuli, helping better explain variance in voxelwise encoding models. Two key methods for determining the unique contribution of each feature space to the variance explained are variance partitioning and the residual method. However, no systematic comparison has been conducted to assess their suitability and properties. To address that gap, this work compares both methods by evaluating them in simulated and real-world experiments and comparing their results. Our findings reveal that both variance partitioning and the residual method can effectively determine the unique variance a feature space explains. However, the residual method requires careful verification of the linear dependence between feature spaces, a step that variance partitioning does not need.

Keywords: Variance Partitioning; Residual Method; Voxelwise Encoding Models; ANOVA

Introduction

Recent studies using naturalistic stimuli in computational cognitive neuroscience have enabled a more diverse analysis of functional magnetic resonance imaging (fMRI) brain recordings (Deniz, Nunez-Elizalde, Huth, & Gallant, 2019; De Heer, Huth, Griffiths, Gallant, & Theunissen, 2017; Gong et al., 2023). When combined with advances in computational modeling (e.g., deep neural networks), this approach allows for the extraction of richer and more diverse feature spaces–quantified representations of stimulus properties that are hypothesized to relate to brain responses as modeled by voxelwise encoding models (VM's) (Lescroart, Stansbury, & Gallant, 2015)). To interpret stimulus representations obtained from models and examine their impact on predicting brain responses, prior studies have proposed two methods: (1) variance partitioning—-a statistical method for estimating the unique and shared variance explained by different feature sets (Borcard, Legendre, & Drapeau, 1992), and (2) the residual approach—-which removes shared information from target features by applying a linear transformation between the source and target feature spaces (Toneva, Mitchell, & Wehbe, 2022). However, no systematic comparison has been conducted between these two methods to assess how effectively they disentangle the unique and shared variance explained by different feature spaces in the brain. This raises the question whether both methods are equally effective in explaining which aspects of model-derived features uniquely contribute to brain responses?

To address this question, in this work, we systematically investigate and compare the two methods in both simulated settings and naturalistic brain recordings, depicted in Figure 1. In simulated settings, we find that both methods can effectively estimate the unique variance explained by a feature space; however, the residual method requires careful verification of linear dependence between feature spaces–a step not needed in variance partitioning. Further analysis of ridge and ordinary least squares (OLS) techniques within the residual method reveals that ridge regression tends to perform less optimally than ordinary least squares in experimental scenarios. In naturalistic story reading/listening experiments, analyzing multiple feature spaces—such as high-level semantic features versus low-level features (e.g., motion energy and number of letters)—reveals important differences in how spurious correlations with brain responses are handled. For the letter-based feature space, both methods successfully reduce spurious correlations. However, when motion energy is used as the feature space, the residual approach is less effective than variance partitioning in isolating unique contributions, suggesting that variance partitioning may be more robust to the collinearity between low-level features and other representational spaces.

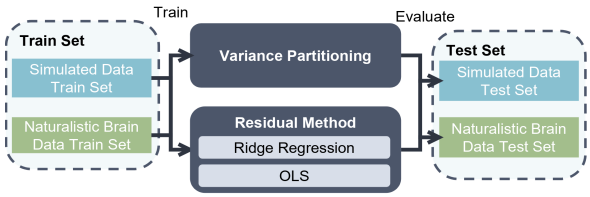

Figure 1: Variance partitioning and the residual method using ridge regression and OLS for regressing features were fit using simulated and experimental test sets. Evaluation was performed on held-out test sets.

Methods

Dataset. To compare variance partitioning and residual methods, we perform analyses on both simulated data and fMRI recordings. Simulations were designed to control the amount of shared and unique variance between two feature spaces and their impact on the simulated brain responses. The experimental data was collected by Deniz et al. (2019) and contained feature spaces and fMRI recordings of subjects reading narrative stories.

Feature Spaces. The feature spaces used in this work include a semantic feature space, which encoded the correlation of each word with the basic English words according to Wikipedia (985 features; Huth, De Heer, Griffiths, Theunissen, and Gallant (2016)). A letters feature space (26 features; Deniz et al. (2019)), which counted the letters presented to the subject at any given time, and a motion energy feature space (4028 features; Deniz et al. (2019)), which encoded the visual motion of letters seen by the subjects.

Variance Partitioning. To estimate the unique and shared variance explained by two feature spaces (X_1 and X_2), we estimate the variance explained by them separately and jointly. The variance explained separately was estimated by fitting two ridge regression models (Hoerl & Kennard, 1970), one for each feature space. The variance explained by both feature spaces jointly was determined using banded ridge regression (Nunez-Elizalde, Huth, & Gallant, 2019), which accounts

for different regularization strengths for each feature space. The variance explained uniquely by each feature space is computed using set theory:

$$R^2_{shared} = R^2_{X_1} + R^2_{X_2} - R^2_{X_1 \text{and} X_2} \quad (1)$$

$$R^2_{X_i unique} = R^2_{X_i} - R^2_{shared}, i \in \{1,2\} \quad (2)$$

Residual Method on Feature Spaces. The residual method aims to remove information encoded by both feature spaces (X_1 and X_2) using a linear transformation (S. Oota, Gupta, & Toneva, 2023). The residuals of this linear transformation f, represent the information uniquely encoded in the regressed feature space: $X'_1 = X_1 - \hat{X}_1$, where $\hat{X}_1 = \hat{f}(X_2)$. Typically f is found using ridge regression (S. Oota et al., 2023; Toneva et al., 2022; S. R. Oota, Çelik, Deniz, & Toneva, 2024). However, regularization may lead to the incomplete removal of shared information, which is not desirable. To explore this issue, our comparisons also include an OLS version of the residual method which removes all (linearly) shared information. Now, a separate linear transformation g can be used to capture the variance in any given voxel Y uniquely explained by X_1 using the residual feature space X'_1: $\hat{g}(X'_1) \approx Y$.

Results

Simulated and experimental data were used to evaluate variance partitioning and the residual method.

Analysis on Simulated Data. Figure 2 shows the comparison between two methods on simulated data. These simulations suggest that variance partitioning is more robust than both residual methods, with the residual method using OLS and ridge regression performing differently depending on the condition.

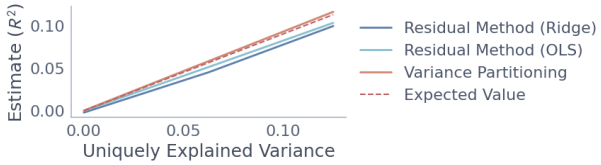

Figure 2: Simulated data comparing two feature spaces with varying unique variance. Variance partitioning slightly overestimates the ground truth, while residual methods tend to underestimate it. The OLS method proved to be less erroneous compared to the ridge method.

Analysis on Brain Dataset. In the experimental data, the semantic feature space accurately predicted brain regions previously associated with semantics (prefrontal cortex, temporoparietal junction) and also regions unrelated to semantics (visual cortex for reading) . To predict brain activity that is only related to semantics and remove spurious correlations, a feature space unrelated to semantic processing (e.g. motion energy) can be used in addition to the semantics feature space. Using the letters feature space, all three methods (Residual method with Ridge regression, Residual method with OLS, Variance Partitioning) reduce the spurious correlations (i.e. high prediction accuracy values of semantic features in visual cortex). Because activity in the visual cortex is no longer predicted by semantic features, activity unrelated to semantic processing was successfully explained away by the letters feature space, as seen in Figure 3. However, when applying the three methods using the motion energy feature space we see that the residual method using both OLS and ridge regression does not reduce spurious correlations as well as variance partitioning. The residual method using ridge regression especially does not reduce spurious correlations in the visual cortex, incorrectly suggesting that the high prediction accuracy of semantic features in visual cortex cannot be explained away by motion energy features.

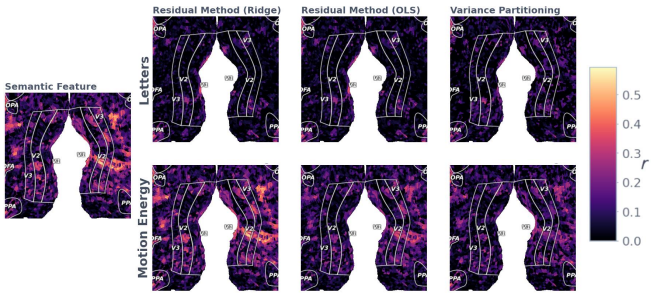

Figure 3: Prediction accuracy of semantic feature space in visual cortex before and after applying variance partitioning and the residual method (lower is better *a priori*). Across methods, semantic features explain little unique variance beyond letters. For motion energy, the ridge residual method attributes the most variance, while variance partitioning attributes the least.

Discussion & Conclusion

In this study, we systematically investigate how variance partitioning and the residual method can be used to determine the variance uniquely explained by a feature space in relation to other feature spaces. Using the residual method requires careful verification of linear predictability between feature spaces. Among the two variants, the residual method with OLS proves more effective at removing shared information between feature spaces compared to its counterpart using ridge regression. More detailed analysis using more than two feature spaces, which is how the methods are usually applied (Deniz et al., 2019; De Heer et al., 2017), is still necessary to fully evaluate variance partitioning and the residual method. The simulated data can also be further improved by confounding the linear relationship between feature spaces to assess the assumptions of the residual method.

Acknowledgement

We thank the reviewers for their time, support, and valuable feedback. This work was funded by grants from the German Federal Ministry of Education and Research (BMBF; Grant no. 01GQ1906) and the European Research Council (ERC; Grant no. 101042567).

References

Borcard, D., Legendre, P., & Drapeau, P. (1992). Partialling out the spatial component of ecological variation. *Ecology*, *73*(3), 1045–1055.

De Heer, W. A., Huth, A. G., Griffiths, T. L., Gallant, J. L., & Theunissen, F. E. (2017). The hierarchical cortical organization of human speech processing. *Journal of Neuroscience*, *37*(27), 6539–6557.

Deniz, F., Nunez-Elizalde, A. O., Huth, A. G., & Gallant, J. L. (2019). The representation of semantic information across human cerebral cortex during listening versus reading is invariant to stimulus modality. *Journal of Neuroscience*, *39*(39), 7722–7736.

Gong, X. L., Huth, A. G., Deniz, F., Johnson, K., Gallant, J. L., & Theunissen, F. E. (2023). Phonemic segmentation of narrative speech in human cerebral cortex. *Nature communications*, *14*(1), 4309.

Hoerl, A. E., & Kennard, R. W. (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, *12*(1), 55–67.

Huth, A. G., De Heer, W. A., Griffiths, T. L., Theunissen, F. E., & Gallant, J. L. (2016). Natural speech reveals the semantic maps that tile human cerebral cortex. *Nature*, *532*(7600), 453–458.

Lescroart, M. D., Stansbury, D. E., & Gallant, J. L. (2015). Fourier power, subjective distance, and object categories all provide plausible models of bold responses in scene-selective visual areas. *Frontiers in computational neuroscience*, *9*, 135.

Nunez-Elizalde, A. O., Huth, A. G., & Gallant, J. L. (2019). Voxelwise encoding models with non-spherical multivariate normal priors. *Neuroimage*, *197*, 482–492.

Oota, S., Gupta, M., & Toneva, M. (2023). Joint processing of linguistic properties in brains and language models. *Advances in Neural Information Processing Systems*, *36*, 18001–18014.

Oota, S. R., Çelik, E., Deniz, F., & Toneva, M. (2024). Speech language models lack important brain-relevant semantics. In *Proceedings of the 62nd annual meeting of the association for computational linguistics (volume 1: Long papers)* (pp. 8503–8528). Association for Computational Linguistics.

Toneva, M., Mitchell, T. M., & Wehbe, L. (2022). Combining computational controls with natural text reveals aspects of meaning composition. *Nature computational science*, *2*(11), 745–757.