Quantitatively comparing predictive models with the Partial Information Decomposition

C Daube[1], BL Giordano[2], PG Schyns[1], RAA Ince[1]
christoph.daube@gmail.com, brungio@gmail.com, {robin.ince,philippe.schyns}@glasgow.ac.uk
[1]Institute of Neuroscience and Psychology, University of Glasgow, Glasgow, UK
[2]Institut de Neurosciences de la Timone UMR 7289 CNRS and Aix-Marseille Universite, Marseille, France

Abstract:

There is increasing focus in cognitive and computational neuroscience on the use of encoding and decoding models to gain insight into cognitive processing. Frequently, encoding models are fit to a number of different features sets, and the out-of-sample predictive performance of the resulting models is compared. However, to gain the maximum benefit from this modelling, we need to go beyond simply ranking model performance in terms of absolute predictive power. We also need to directly compare and relate the predictions between models, to gain insight into which models are predicting common vs unique aspects of the neural response. The Partial Information Decomposition (PID) provides a principled theoretical framework to address this question, as it decomposes the total predictive performance of two models into redundant (overlapping), unique, and synergistic parts. We show that like classical information theoretic quantities, variance decomposition approaches conflate synergy and redundancy and so could provide a misleading view of the unique predictive power of a model. We also suggest how the use of encoding models and PID can help interpret decoding models.

Keywords: model comparison, variance decomposition, partial information decomposition, information theory

Comparing model predictions

Existing Approaches A number of studies have used variance-based approaches to partition unique and common variance predicted by different models. While these rely on common approaches to comparing variance (Seibold and McPhee, 1979), they differ in how they are applied. One approach involves comparing R^2 values directly from predictive models fit on combinations of feature spaces (Heer et al., 2017; Lescroart et al., 2015). However, this suffers from the complication that the model fitting procedure, including regularization, might interact with the different feature spaces when they are considered separately vs jointly. An alternative approach is to perform the variance partitioning as a "second-level" analysis, applied to the low-dimensional predictions of the individual models. The second-level approach has the advantage that any modelling approach can be used for the two considered feature models – they are not restricted to linear regression. The approach can also be applied to generalized linear models, neural network model predictions, representational geometries (Hebart et al., 2018), etc.

Partial Information Decomposition The PID provides a framework to decompose the information conveyed about a target by a set of sources, into that which is unique to each, redundant (or common) to each subset or synergistic between each subset (Ince, 2017; Park et al., 2018; Williams and Beer, 2010). Here we consider the predictions of two models on a hold-out test set as the two sources, and the corresponding observed neural data as the target (Daube et al., 2019). Note that here the predictions of each model are directly compared as a second-level analysis after the individual model fitting without using a separate joint model. Redundancy quantifies the common predictive power between the two models. Unique information provides the unique predictive power of each – i.e. quantifies the occasions where one model would correctly predict a sample while the other wouldn't. Synergy quantifies that the sample-by-sample relationship between the two predictions encapsulates extra predictive information about the data that is not captured by either model alone, and therefore suggests that a joint model combining the features should be investigated.

The PID satisfies the following relationships (Williams and Beer, 2010) for the joint mutual information (jMI), conditional mutual information (CMI) and co-information (co-I) respectively (Cover and Thomas, 1991; Ince et al., 2017), where P_1, P_2 represent the predictions from two different models and D represents the held-out data they are predicting.

$$I(P_1, P_2; D) = \mathrm{Red} + \mathrm{Unq}(P_1) + \mathrm{Unq}(P_2) + \mathrm{Syn}$$
$$I(P_1; D|P_2) = \mathrm{Unq}(P_2) + \mathrm{Syn}$$
$$I(P_1; P_2; D) = \mathrm{Red} - \mathrm{Syn}$$

This demonstrates that CMI should not be interpreted as the unique contribution of one source, because it includes also the synergy between the sources. Similarly, co-I combines redundancy and synergy quantifying the net interaction effect.

Results

Gaussian system Figure 1 shows results for a tri-variate Gaussian system where each model prediction has a correlation of 0.5 with the hold-out target data. Variance and information decompositions are plotted as a function of the correlation between the model predictions. i.e. p=1 means both models predict exactly the same value on every trial, p=0 means the model predictions are uncorrelated. Note that commonality looks very similar to the information theoretic co-I, which conflates redundancy and synergy (blue lines). Unique variance (semi-partial correlation) looks very similar to CMI (yellow lines), which again conflates unique information and synergy. Therefore, this suggests that commonality analysis is also unable to separate redundant and synergistic effects, and therefore cannot accurately quantify the unique predictive power of a model. Further, this suggests that negative common variance terms, may, like negative co-I, reflect synergistic relationships. In commonality analysis these negative commonality terms are often taken to indicate the presence of a mediating 'suppressor' variable (Ray-Mukherjee et al., 2014), although in neuroimaging some authors employ ad-hoc corrections to them (Heer et al., 2017).

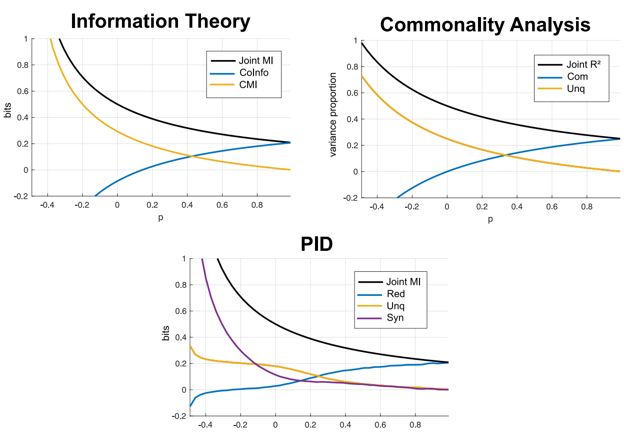

Figure 1: Classical information theoretic measures, Commonality analysis and PID results for a Gaussian system with equal model prediction performance ($\rho = 0.5$) as a function of the correlation of the model predictions p. Due to the symmetry of the system there is equal unique in each predictor.

Auditory MEG Encoding Models During Speech Figure 2 shows results of comparisons of models predicting auditory cortex responses to speech with ridge regression over a range of non-linear auditory feature sets. Close to 100% redundancy between predictions based on features consisting of the spectrogram and its rectified derivative (Sg&Deriv) and the benchmark oracle model (Kriegeskorte and Douglas, 2018) based on phonetic features (Di Liberto et al., 2015) was observed. This indicates that the tested model captures all the predictive power of the benchmark model. The unique information shows that while the benchmark model has no unique predictive power, the tested Sg&Deriv model does have unique predictive information.

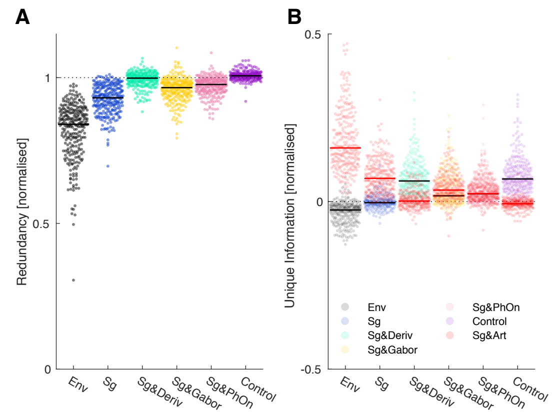

Figure 2: Comparing MEG encoding models with the PID. A range of models based on different features spaces are compared to a benchmark model. Left shows redundancy (normalized to predictive information of benchmark model). Right shows unique information in tested model and benchmark model (red). Data points are from different cross-validation outer folds and participants. Data and analysis from Daube et al., 2019.

Interpreting decoding models with the PID

While the interpretational differences between encoding and decoding models have been described (Ahissar et al., 1997; Holdgraf et al., 2017; Kriegeskorte and Douglas, 2018; Naselaris et al., 2011; Weichwald et al., 2015), it remains tempting to interpret strong decoding of a high-level stimulus feature (e.g. phoneme class) as a signature of a processing stage that operates on that feature (e.g. pre-lexical abstraction) (Khalighinejad et al., 2017). However, the decoding performance could occur as a side-effect of low-level differences between the decoded high-level features. Using PID model comparison this can be tested explicitly, by applying PID with two sources: measured data, and hold-out predictions of a low-level feature model. If it is possible to decode the high-level feature from the predictions of the low-level feature model, and, crucially, this decoding is redundant with the decoding performed on the measured data, then this is strong evidence that the high-level decoding is an epiphenomenon of a low-level sensory encoding (Daube et al., 2019).

Conclusions

The partial information decomposition has received much recent interest, as it invites a reinterpretation of classical information theoretic quantities. In particular, co-information and conditional mutual information need to be reconsidered in light of the fact that these measures also quantify synergistic effects. In model comparisons, a zero value of co-information (similarly common variance) thus does not mean there is no overlapping prediction between the models, because redundant predictive effects could be cancelled out by different but equally strong synergistic effects. A high CMI does not mean one model conveys unique predictive information, because it may reflect synergy. PID provides a promising approach to perform systematic model comparison while accounting for these potential confounds, which also affect variance partitioning methods. PID can be applied as a second-level analysis to compare and interpret predictive encoding and decoding models, as well as models of representational similarity.

Acknowledgments

RI was supported by the Wellcome Trust [214120]. PGS was supported by the Wellcome Trust [107802] and the MURI/EPSRC [172046-01].

References

Ahissar, E., Haidarliu, S., and Zacksenhouse, M. (1997). Decoding temporally encoded sensory input by cortical oscillations and thalamic phase comparators. Proceedings of the National Academy of Sciences *94*, 11633.

Cover, T.M., and Thomas, J.A. (1991). Elements of information theory (Wiley New York).

Daube, C., Ince, R.A.A., and Gross, J. (2019). Simple Acoustic Features Can Explain Phoneme-Based Predictions of Cortical Responses to Speech. Current Biology *0*.

Di Liberto, G.M., O'Sullivan, J.A., and Lalor, E.C. (2015). Low-Frequency Cortical Entrainment to Speech Reflects Phoneme-Level Processing. Current Biology *25*, 2457–2465.

Hebart, M.N., Bankson, B.B., Harel, A., Baker, C.I., and Cichy, R.M. (2018). The representational dynamics of task and object processing in humans. ELife Sciences *7*, e32816.

Heer, W.A. de, Huth, A.G., Griffiths, T.L., Gallant, J.L., and Theunissen, F.E. (2017). The hierarchical cortical organization of human speech processing. J. Neurosci. 3267–16.

Holdgraf, C.R., Rieger, J.W., Micheli, C., Martin, S., Knight, R.T., and Theunissen, F.E. (2017). Encoding and Decoding Models in Cognitive Electrophysiology. Front. Syst. Neurosci. *11*.

Ince, R.A.A. (2017). Measuring Multivariate Redundant Information with Pointwise Common Change in Surprisal. Entropy *19*, 318.

Ince, R.A.A., Giordano, B.L., Kayser, C., Rousselet, G.A., Gross, J., and Schyns, P.G. (2017). A statistical framework for neuroimaging data analysis based on mutual information estimated via a gaussian copula. Hum. Brain Mapp. *38*, 1541–1573.

Khalighinejad, B., Silva, G.C. da, and Mesgarani, N. (2017). Dynamic Encoding of Acoustic Features in Neural Responses to Continuous Speech. J. Neurosci. 2383–16.

Kriegeskorte, N., and Douglas, P.K. (2018). Cognitive computational neuroscience. Nature Neuroscience *21*, 1148.

Lescroart, M.D., Stansbury, D.E., and Gallant, J.L. (2015). Fourier power, subjective distance, and object categories all provide plausible models of BOLD responses in scene-selective visual areas. Front. Comput. Neurosci. *9*.

Naselaris, T., Kay, K.N., Nishimoto, S., and Gallant, J.L. (2011). Encoding and decoding in fMRI. NeuroImage *56*, 400–410.

Park, H., Ince, R.A.A., Schyns, P.G., Thut, G., and Gross, J. (2018). Representational interactions during audiovisual speech entrainment: Redundancy in left posterior superior temporal gyrus and synergy in left motor cortex. PLOS Biology *16*, e2006558.

Ray-Mukherjee, J., Nimon, K., Mukherjee, S., Morris, D.W., Slotow, R., and Hamer, M. (2014). Using commonality analysis in multiple regressions: a tool to decompose regression effects in the face of multicollinearity. Methods Ecol Evol *5*, 320–328.

Seibold, D.R., and McPhee, R.D. (1979). Commonality Analysis: A Method for Decomposing Explained Variance in Multiple Regression Analyses. Human Communication Research *5*, 355–365.

Weichwald, S., Meyer, T., Özdenizci, O., Schölkopf, B., Ball, T., and Grosse-Wentrup, M. (2015). Causal interpretation rules for encoding and decoding models in neuroimaging. NeuroImage *110*, 48–59.

Williams, P.L., and Beer, R.D. (2010). Nonnegative Decomposition of Multivariate Information. ArXiv:1004.2515 [Math-Ph, Physics:Physics, q-Bio].