Finding Nonlinear Relationships in fMRI Time Series with Symbolic Regression

James Alexander Hughes
Computer Science, Brain and Mind Institute
University of Western Ontario
1151 Richmond St.
London, Ontario, Canada N6A 3K7
jhughe54@uwo.ca

Mark Daley
Computer Science, Brain and Mind Institute
University of Western Ontario
1151 Richmond St.
London, Ontario, Canada N6A 3K7
mdaley2@uwo.ca

ABSTRACT

The brain is an intrinsically nonlinear system, yet the dominant methods used to generate network models of functional connectivity from fMRI data use linear methods. Although these approaches have been used successfully, they are limited in that they can find only linear relations within a system we know to be nonlinear.

This study employs a highly specialized genetic programming system which incorporates multiple enhancements to perform symbolic regression, a type of regression analysis that searches for declarative mathematical expressions to describe relationships in observed data.

Publicly available fMRI data from the Human Connectome Project were segmented into meaningful regions of interest and highly nonlinear mathematical expressions describing functional connectivity were generated. These nonlinear expressions exceed the explanatory power of traditional linear models and allow for more accurate investigation of the underlying physiological connectivities.

Keywords

Symbolic regression; Computational neuroscience; Functional magnetic resonance imaging; Nonlinear relationships

1. INTRODUCTION

Literature in the field of neuroscience explicitly acknowledges the existence of nonlinear relationships in brain function [1, 3], but it is common to treat them as a footnote or ignore them altogether [2, 3]. Linear tools, such as the General Linear Model (GLM) or the Pearson product-moment coefficient are used, almost exclusively, to model functional magnetic resonance imaging (fMRI) time series. Despite this, neuroscientific studies are able to make contributions with limited linear model [1]; however, it would ultimately be improper to use linear methods to observe what we *know* to be nonlinear phenomenon as it lacks the power to truly model the underlying processes. It is not surprising that the nonlinear relationships are ignored; discovering underlying nonlinearities is an exceptionally non-trivial task, especially when working with large amounts of high-dimensional data.

In this work Genetic Programming (GP) is implemented to automate the discovery of minimal and interpretable network relationships in the behavior of a system for which we can observe only time series derived from a network's nodes: task based fMRI time series data. No prior knowledge or assumptions are applied to the system, such as linearity or how the system interacts with itself.

GECCO '16 July 20-24, 2016, Denver, CO, USA

ACM ISBN 978-1-4503-4323-7/16/07.
DOI: http://dx.doi.org/10.1145/2908961.2909021

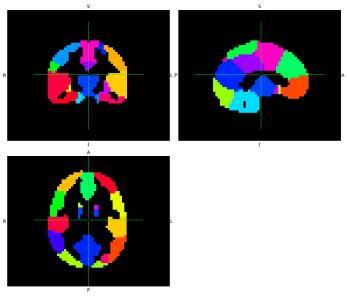

Figure 1: Snapshot of a brain segmented into the 30 ROIs. Each color represents a different region.

2. EXPERIMENTAL METHODS

The task based fMRI time series data selected was of a Motor task and was obtained from the Human Connectome Project, WU-Minn Consortium[1]. This four-dimensional data (three-dimensional brain over time) was collected into *30 spatial regions of interest* (ROIs) (Figure 1) for the time series of *284 time points*, and can be represented as a two-dimensional matrix of *30* columns with *284* rows.

This specific GP implementation is motivated by Schmidt et al.'s work [6], is extremely specialized for symbolic regression, and incorporates modular improvements which significantly increase performance. These improvements including parallel evolution of subpopulations, fitness predictors [5], and an acyclic graph representation [4].

For symbolic regression, it was required to have some value over the time series that evolved expressions fits to. For the purpose of this motor task, *ROI 21* was selected for the left hand side of the equation as it is the ROI that contains the *primary motor cortex*. *100* models for all *507* subjects available were generated.

[1]http://www.humanconnectome.org/

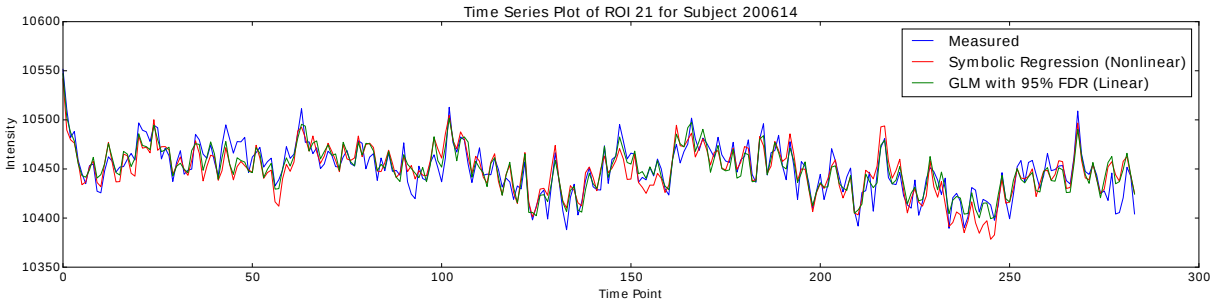

Figure 2: Time series of ROI 21's signal compared to the generated nonlinear and linear models.

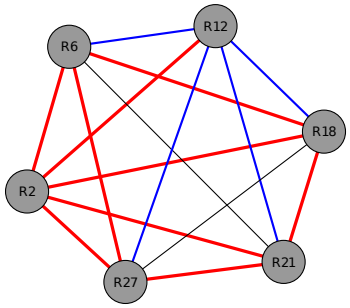

Figure 3: Relationships between ROIs for a single generated *nonlinear model*. Red represents nonlinear relationships *between ROIs*, blue represents nonlinear and linear relationships, and black is strictly linear. This particular example corresponds to the equation: $R21 = R12 - sin(11.97 * (18.30 - R12)) - (0.42 * |(R12 - R18) * R27|)/(R6 - tan(R2))$.

3. RESULTS AND CONCLUSIONS

With Pearson product-moment correlation and false discovery rate (FDR) thresholding (typical linear methods), almost every ROI is linearly related to ROI 21 (on average, 28 ROIs were related to ROI 21 per subject). Performing linear regression with this many ROIs generates models with high degrees of freedom that fit the data well, but provide minimal insight and are difficult to interpret.

Figure 2 shows a time series of one subject's recorded signal alongside two models describing the signal — one found with the nonlinear tool (Figure 3), the other with linear regression after thresholding ROIs with a 95% FDR. The *mean absolute error* over the time series for the top nonlinear models and the thresholded linear models were averaged over all subject. These values were roughly *16.68 (sd = 3.51)* and *11.79 (sd = 1.11)* respectively. Although both models fit the data well, a Mann-Whitney U test (U-test) provides a p-value of $3.08 * 10^{-133}$, which demonstrates that the linear models fit the recorded signal better.

On average, a nonlinear model contained fewer than 4 ROIs (3 when excluding ROI 21). The mean absolute time series error of the linear models generated with the top 4 correlated ROIs — which were typically the same ROIs as those found with GP — was calculated to be approximately *19.16* (sd = *5.08*). A U-test comparing the 4 ROIs models provided a p-value of $8.56 * 10^{-19}$; *the nonlinear models were significantly better*. In fact, it was not until the linear models were given the top *8 ROIs* that there was no more statistical difference. Linear models only performed better than the nonlinear models with 4 ROIs once they received *10 or more ROIs* (U-test p-value of $1.34 * 10^{-3}$); it took at least 10 ROIs for a linear model to fit the recorded signal better than a nonlinear model containing only 4.

When compared to linear models generated with all ROIs available after a typical thresholding technique, nonlinear models, although close, could not fit the signal as well. However, these linear models would typically contain more than 28 ROIs and would be difficult to interpret and provide minimal insight into understanding the underlying processes. Nonlinear models, in contrast, were more succinct and describe nonlinear relationships that would otherwise *not be discovered with conventional tools*. On average, with just 4 ROIs, a nonlinear model could fit the recorded signals better than linear models using 8; even with more information (ROIs), linear models could not describe the data as clearly.

4. ACKNOWLEDGMENTS

This research was supported in part by the Natural Sciences and Engineering Research Council of Canada (NSERC). Computations were enabled by the SciNet HPC Consortium. Data were provided by the Human Connectome Project, WU-Minn Consortium.

5. REFERENCES

[1] R. Buckner and T. Braver. Event-related functional mri. In P. Bandettini and C. Moonen, editors, *Functional MRI*, chapter 36, pages 441–452. Springer-Verlag.

[2] M. Daley. An invitation to the study of brain networks, with some statistical analysis of thresholding techniques. In *Discrete and Topological Models in Molecular Biology*, pages 85–107. Springer, 2014.

[3] N. K. Logothetis. What we can do and what we cannot do with fmri. *Nature*, 453(7197):869–878, 2008.

[4] M. Schmidt and H. Lipson. Comparison of tree and graph encodings as function of problem complexity. In *Proceedings of the 9th annual conference on Genetic and evolutionary computation*, pages 1674–1679. ACM, 2007.

[5] M. D. Schmidt and H. Lipson. Coevolution of fitness predictors. *Evolutionary Computation, IEEE Transactions on*, 12(6):736–749, 2008.

[6] M. D. Schmidt, R. R. Vallabhajosyula, J. W. Jenkins, J. E. Hood, A. S. Soni, J. P. Wikswo, and H. Lipson. Automated refinement and inference of analytical models for metabolic networks. *Physical biology*, 8(5):055011, 2011.